DOCUMENT RESUME

ED 320 967 TM 015 294

AUTHOR Burry, Judith A.; And Others
TITLE Validity and Reliability of Classroom Observations: A Paradox.
PUB DATE 90
NOTE 32p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Classroom Research; Elementary Secondary Education; *Error of Measurement; Methods Research; *Psychometrics; *Reliability; *Research Methodology; *Validity

ABSTRACT

Seven features of valid classroom observations procedures, a five-step measurement schemata, and a systematic classroom observation procedure designed to reduce measurement error are presented. The following features are addressed: (1) the inference of the observation is established a priori and confirmed post hoc; (2) craft knowledge and expertise in pedagogy and teaching is developmental; (3) instrumentation must be built upon a research base, an empirical base, or a combination of the two; (4) observation criteria are generic and content specific; (5) observers are well-trained in the process of observing; (6) observers are well-honed on the meaning of teaching behaviors to be observed; and (7) procedures vary depending upon the inferences that are made. The five steps include observation, recordkeeping, retrieval of observations, analysis, and judgment. A primary issue addressed is the question of the adequacy of traditional measurement procedures for investigating the psychometric properties of classroom observations. Theoretical and empirical evidence illustrates the need for carefully examining traditional alternative methodologies for establishing psychometric properties. Classroom observations may not appear reliable in the true-score sense, but are valid if all seven features are present. Five data tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JUDITH A. BURRY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED320967

VALIDITY AND RELIABILITY OF CLASSROOM OBSERVATIONS: A PARADOX

Judith A. Burry
The University of Alabama

Brad S. Chissom
The University of Alabama

Dale G. Shaw
The University of Northern Colorado

BEST COPY AVAILABLE

TM 015294

Validity and Reliability of Classroom Observations: A Paradox

Judith A. Burry
The University of Alabama

Brad S. Chissom
The University of Alabama

Dale G. Shaw
The University of Northern Colorado

Abstract

The measurement of classroom behavior is a complex task. Teaching has many facets and is context integrated which makes observing classroom behaviors a complicated undertaking. Presented in this paper are: a set of eight assumptions about classroom observations; a five step measurement schemata; and a systematic classroom observation procedure designed to reduce measurement error. A primary issue is the question of the adequacy of traditional measurement procedures for investigating the psychometric properties of classroom observations. Theoretical and empirical evidence illustrate the need for carefully examining traditional and alternative methodologies for establishing psychometric properties. Classroom observations may not appear reliable in the true-score sense, but in fact are valid: hence, a paradox.

Validity and Reliability of Classroom Observations: A Paradox

Traditionally reliability and validity are at the heart of assessment. In practice, instruments undergo a comprehensive analysis for both psychometric properties before a "stamp of approval" is given. Classroom observations are no exception; however, the test trial has just begun on whether our traditional psychometric conventions are doing the job in ascertaining the reliability and validity of teacher classroom observations.

Some measurement experts may argue that our current practices are sufficient for obtaining proof. After all, have not our traditional conventions served us adequately for nearly a century? Have we not also assumed that our procedures for documenting psychometric properties are doing their job? Broadly speaking, then, it can be said that our methods for demonstrating reliability and validity of classroom observation instruments are adequate. To what extent, then, do these procedures work in teacher classroom performance? On the justification of assessing what sort of evidence may we interpret the reliability and validity of classroom observations to be reliable and valid indicators of our teachers' performance? Current literature is void of answers to these questions. In fact there is a paucity of information even questioning whether or not our traditional methods work. Have we measurement experts created a paradox?

The complexity of measuring the classroom observation

The measurement of classroom behavior is a complex task. Teaching behaviors are many faceted and merge with the context of the classroom environment. Teachers and students must interact if learning is to take place which further complicates the

observation process. Another facet of the classroom observation is the observer who brings to the observation task some form of experience, training, and observer bias. Inspecting these classroom observations for psychometric properties involves examining the instrument (set of teaching behaviors), the teacher, the observer, and the interactions within the context of this multifaceted process.

Systems that appraise the competence of classroom teachers emphasize the use of the classroom observation probably the single most important component for any teacher assessment system is the classroom observation. School districts and state departments of education look to teacher evaluation procedures which utilize the classroom observation as a means to strengthen teacher performance. Most however,are not achieving the goal of enhancing teacher performance (Educational Research Service, 1978; Bolton, 1973). Recently, The Personnel Evaluation Standards commented that " ... educational institutions have often been ineffective in carrying out their personnel evaluation responsibilities (1988, p.6)."

The purpose of this paper is to investigate potential solutions to these problems which in part lie in the understanding of how our traditional measurement properties of reliability and validity apply to performance assessment. An empirical example is presented to illustrate an innovative approach demonstrating reliability and validity of a classroom context classroom observation. This example also serves to illustrate the need to continue exploring old and new methods for obtaining psychometric evidence of quality.

Underlying Features for Classroom Observations

How do we know what a measure of classroom behavior means if different responses are illictited by alternative teaching styles, strategies, context, teaching levels, and content areas? On what proof may we justify using an established measure to license or certify teachers? In what ways do we interpret the significance of the total score? Should classroom observation measures be used for the purpose of hiring, firing, promoting teachers, assisting teachers in the manner proposed? Ultimately, what evidence signifies the functional worth of the classroom measure? These questions relate to the omnibus psychometric concern of validity. Messick states: "... the key issues of test validity are the meaning, relevance, and utility of scores, the important or value implications of scores as a basis for action, and the functional worth of scores in terms of the social consequences of their use" (1989, p. 5). More conventionally, the Joint technical standards for educational and psychological testing (APA, AERA, NCME, 1985) states: "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness, of the specific inferences made from test scores... validity always refers to the degree to which that evidence supports the inferences that are made from test scores." It is fitting then to begin this discussion of the reliability and validity of classroom observations with the definition of validity. In order to bring meaning to classroom observations, we will discuss certain features pertaining to the observations. This, will provide us with a context in which to consider the psychometric issues.

Table 1. Features of Valid Classroom Observation Procedures

1. The observation inference is established apriori and confirmed post hoc.
2. Craft knowledge (expertise in pedagogy) in teaching is developmental.
3. Instrumentation must be built upon a research base, empirical base, or a combination of the two.
4. Observation criteria are both generic and content specific.
5. Observers are well-trained in the process of observating.
6. Observers are well-honed on the meaning(s) of the teaching behaviors to be observed.
7. Procedures must vary depending upon the inference(s) that are made.

The Inference of the Observation is Established Apriori and confirmed post hoc

Since the validity of the classroom observation is the omnibus issue the inference of the observation is established before the instrument is written and confirmed after the instrument has been used for its intended purpose. If the instrument is used for assessment purposes, then both reliability and validity are critical issues. If the instrument is used solely for professional development, then validity is probably the only psychometric consideration. The other aspects of inference are included in the remainder of these assumptions.

Craft Knowledge, Expertise in Pedagogy, in Teaching is Developmental

Scientific knowledge of the the developmental patterns in teaching expertise is short on scientific research. Berliner approaches teaching in a developmental mien as a theory of skill learning with the following stages: 1) Novice, 2) Advanced Beginner, 3) Competent, 4) Proficient, and 5) Expert. He comments: "There are differences in the ways that teachers at various levels of experience and expertise interpret classroom phenomena. (1988, p. 44)"

Berliner describes the Novice Stage as one in which the familiar is ascertained, labeled and learned, and a collection of context-free rules is obtained. Characteristics of the Advanced Beginner are: the teacher is able to depict similarities across contexts, understands episodic knowledge, and uses strategic knowledge - understands when to ignore, break or follow rules-and incorporates context as a guide for teaching behavior. Competent teachers make conscious choices about what they do, set priorities, and plan. A proficient teacher uses intuition and know-how. An expert teacher grasps situations with an intuitive sense without being analytical or deliberative. His/her teaching performance is fluid and gives the appearance of being effortless.

Instrumentation Must be Built Upon a Research Base, or an Empirical Base, or a Combination of the Two

For more than 15 years researchers have designed instruments based upon the process-product literature (Burry, Poggio, & Glasnapp, 1989). Effective teaching behaviors are correlated with measures of student achievement as measured by standardized tests. (See, for example, Brophy & Good, 1986; Denham & Lieberman, 1980; Dunkin & Biddle, 1974.) Underlying this research is the premise that teaching is primarily a linear activity, wherein teaching behaviors are considered causes (process) and pupil learning is viewed as an effect (product).

Lee Shulman and colleagues (Suzanne Willson, Gary Sykes, Tom Bird, Angelo Collins, Linda Vavaros, among others) with the Teacher Assessment Project at Stanford University have taken this stance: "...the elegance of theory must be combined with the wisdom of practice." (Leinhard, 1990, pp. 18-19). These researchers may be considered authors of the interpretative paradigm.

These two paradigms represent the current state of the art in teacher assessment (Cochran-Smith & Lytle, 1990). In assumption two either or both of these paradigms are sufficient for providing the foundation for classroom observation instrumentation.

Observation criteria, when appropriate, are both generic and content specific

Two beliefs exist about the generic and specific aspects of teaching behaviors. For the most part, state used instruments for assessing first year and career ladder teachers include generic behaviors. One set of behaviors is used for doing classroom observations for all grade levels, content areas, and levels of expertise. However, problems exist associated with this notion in that different certification areas view these behaviors as meaningful in different rank orders (Poggio, Glasnapp, & Burry, 1989). Shulman (1988) suggests that teaching is a highly complex context-specific interactive activity in which variations across classrooms, schools, and communities are acutely important. Burry and Shaw (1988) found that teaching behaviors have the same dimensions, but are viewed differently by teachers and administrators observing teachers.

Two opposing beliefs are subsumed withing this assumption that addresses teaching behaviors as being both generic and specific. In time, behaviors that are generic for some pedagogical level of expertise will be delineated as will those that are content oriented. Until then, we assume both the generic and specific aspects of teaching behaviors are appropriate for this discussion.

Observers are well-trained in the process of observing and Observers are well-honed on the meaning(s) of the teaching behaviors to be observed.

Observers must have a proficient knowledge of the behaviors to be observed. The validity of the observation is contingent on the observer's understanding of the teaching

behaviors (Evertson & Burry, 1989). Observers must also have a high degree of skill in the observation procedures particular to the instrumentation. Otherwise the validity of the observation is destroyed.

The Validity of the Observation

In examining the validity of a context classroom observation (Burry, Chissom, & Iran-Nejad, 1989), equated the process with the overlapping of three camera lenses. Three camera lenses are focused simultaneously on the classroom observation in order for the observation to be valid. The lenses address three domains: pedagogy, learning, and measurement. The foundation for this discussion is found in features one and two. Learning, on the part of the teacher, is detailed in feature one and for observers feature five and six. Measurement brings the other two lens into focus and addresses assumption one. If all seven of these features are present, we can say "the observation is valid."

Five Critical Measurement Steps

In the realm of classroom observations it is paramount that the discussion of measurement not be separated from the phenomena of the craft knowledge of teaching feature two). Substantive knowledge is a precursor to obtaining information on any teaching attribute. In a recent chapter in the Handbook of research on teaching (Shavelson, Webb, & Burnstein, 1988) address the psychometric aspects of only two types of data collection due to the complexity of the nature of teacher assessment. They focus on only the classroom observation and selected aspects of measuring teacher's

cognitive processing. In conclusion they write " research on teaching is a psychometrician's nirvana and nemesis at the same time. As the research in this areas progresses at an explosive rate, measurement issues abound. There is practically no limit to the measurement problems in need of solution. (p.86)." We shall use this quote as a caveat in our inquiries.

One way to control for inherent measurement problems is to standardize the procedures for collecting data. To obtain, analyze, and provide data for decision making we must have a consistent strategy that will enable us to arrive at accurate and sound decisions. In order to control for potential external and internal factors which contribute to measurement error, a five step model was designed (Burry, Poggio, & Glasnapp, 1988) based upon Landy and Farr's (1983) model of performance assessment. These five steps can be applied to any kind of instrumentation used in teacher performance assessment which, of course, includes the classroom observation. Table 2 lists the five steps.

Table 2. The Five Observation Steps

1. **Observe** using an appropriate instrument.
2. **Record** objective and accurate data.
3. **Retrieve** the observation using the record.
4. **Analyze/Assist** the observation by comparing the record from the data source to a standardized criterion. Assistance can be easily integrated at this step so as to avoid a reanalysis of the data. When the observation is compared to a criterion it is easy to see what the teacher need to do to improve.
5. **Judge** the observation using the information from the analysis step.

This five step observation schemata was designed to minimize measurement error by controlling for potential external and internal factors at all aspects of the classroom observation process. All five steps must be used regardless of the purpose of the observation whether it be research, purely assistance (professional growth), purely assessment, and or a combination of professional growth and assessment.

Traditionally the observer has carried a check list into the classroom, observed the teacher, and rated the teacher on a Likert-type scale. In this case steps one and five occur almost simultaneously. The observer leaves the classroom with a completed observation in that he/she has checked off some characteristic and a rating of the teaching behaviors. A written document exists, but only in the form of a summative rating. What is missing is the context and a record of the classroom proceedings.

Performing these five steps is critical to the validity of the classroom observation. Making judgments simultaneously while observing and recording information greatly increases the opportunity for error. When recorded information is not used in the classroom observation process, it is analogous to taking a picture without film. The recorded information serves as the basis for the analysis/assistance and ultimately the judgment. Performing these five ordered steps serves to increase the objectivity, precision, reliability and validity of the classroom observation. There is no written evidence of the observation, which means that unless the events were video tapped there is no way of retrieving and analyzing the observation. If professional recommendations are to be made, there is no bases for comparison to a criteria. Another problem that exists with this kind of classroom observation is that rater and ratee biases can be operating sometimes at their maximum. Finally, currently used procedures for establishing reliability and validity of the classroom observation are limited.

Reliability of the Five Step Classroom Observation procedures

Classical and modern measurement theory provide a number of insights into what we may do to ensure or enhance the reliability of any process for assessing teacher performance. Although the theories by no means offer guarantees of instrument quality, they do present us with specific suggestions for writing or revising instruments which will have good psychometric properties.

Obviously, the primary concern in any assessment process is that it be valid. Measurement theory, however, tells us that a process without reliability cannot be valid (Crocker & Algina, 1986). Further, high reliability indicates only that random error has been controlled, and thus by itself does not guarantee validity. Non-random error such

as observer bias for/against the observee actually is treated as part of the true score component in the classical true score model X (observed score) = T (true score) + E (random error)(Magnussom, 1966). Yet regardless of whether error is random (affects reliability), or not random (affects validity and is not captured in reliability indices), it is a worthy goal to attempt to reduce sources of either type of error in an assessment procedure.

At each step of the five-step observation schemata, there are specific guidelines which measurement theory tells will reduce or likely reduce error and enhance reliability. Because the effect of an error can magnify itself as the steps progress, it is important that the error be maximally controlled at each step.

In the first step, that of observing specific indicators of behavior, the reliability of these observations is critical. Obviously each behavior to be observed must be succinctly and unambiguously defined, examples constituting the behavior or absence of it clearly understood, and the methods for precisely recording it clearly understood by all persons serving as observers.

Not only must observers be trained in what to look for, they must also be trained to avoid typical rating errors such as the tendency to be to lenient and the tendency to rate from an overall impression (halo-effect) (Ary, etal., 1990). There should be, if resources permit, more than one observer, with observations being combined in some manner. If two observers are used then a third rater might be called in if there are substantial discrepancies among their observations. Observers should periodically rate some fixed standard or all be present to rate the same situation for the purpose of addressing equivalency of observations.

At step two, where observations are recorded for a permanent, retrievable record, measurement theory tells us that, where possible, to provide the greatest precision of measurement. Behaviors being observed should be scaled on at least a four-interval continuum, and preferably five to seven intervals (Shaw, etal, 1987). Reliability is dramatically reduced when behaviors are recorded with a Y/N response instead of calibrated over a five-interval scale. Also responses must be reproducible by the observer.

Step three calls for the retrieval of the observational record. Observers must have as their focus the production of a written record of observations that is readable, unambiguous, and succinct. Their training should emphasize that their primary mission is to produce this accurately prepared record of observations. Theoretically, someone else could come behind them and do steps four and five, because all salient information to be used in the assessment is clearly recorded and retrievable from the document.

At step four, matters become somewhat more complicated. If one believes that teacher behavior items constitute a unidimensional construct, then measurement theory suggests combining item observations into a single composite score and submitting this score (Allen & Yen, 198 ; and Crocker and Algina, 1986). If one believes that the items really define difference facets of teaching, then subscale scores should be obtained if the facets are independent. If they are related, then only one composite score is interpretable. Of course, feedback on each teaching behavior can be useful to the observee. Might we not do both? Should we not provide specific item feedback to teachers as well as a composite score analysis?

The fifth step, reliability of judgment is best served using either Hoyt's Analysis of Variance or Generalizability Theory (Crocker & Algina, 1986). Judgement can be rendered at the behavior level or composite score level, no subjectivity must enter the process.

Establishing Reliability for Classroom Observations

Reliability is critical to classroom observations that are used for licensing or certification. Teaching behavior items are standardized and observers are trained on obervation procedures as well as the meaning of the behavior items. The issue of reliability has multiple meanings. Are the behavior items consistently measuring teaching behavior? Is there an agreement between two raters on one classroom observation? Is there an agreement for one rater across multiple observations for one teacher? Is there an agreement between raters across multiple teachers on one set of behaviors, etc. Reliability in the true score sense is different from measures of interrater reliability used by many assessors.

Our traditional measurement theory includes a theoretical premise for obtaining a "true score" from an observed score. The theory states that every observed score is the sum of two components, a true score and random error score. Our measurement community has offered through generalizability (G) theory, an analysis of variance application, an extension of true score theory. G theory allows us to partitision our sums of squares into variance components for each facet of our observation: behavior items, subject, and raters, the interactions, and the error term(s).

The process of isolating random error is considerably more complicated, because, according to our Five Step Observation Schemata, there are five places were random

error can occur. The psychometric property of reliability is complicated beyond one's imagination when all five steps of the observation schemata are in place. A major question to be asked is what is reliability of a performance observation? Traditionally, we have used interrater reliability to determine an aspect of reliability for a classroom observation. In the true score sense, interrater or observer reliability is not reliability. It is, at best, a measure of the degree of relationship between two observers for a given set of measures (Shavelson, Webb, & Burstein, 1986; Crocker & Algina, 1986). High observer agreement is not reliability; however, high observer agreement is necessary prescription for high reliability.

Generalizability (G) theory, a study of the universe of generalization, permits the researcher to estimate multiple sources of variation in observations simultaneously. G theory is considered an extension and liberalization of classical theory (Feldt & Brennan, 1989). The foundation of generalizability theory is an analysis of variance incorporating as facets multiple sources of variation such as: (1) teachers, (2) observers, (3) behaviors, and (4) the sources of error and interaction between the facets (Shavelson, Webb, & Burnstein, 1986; Feld & Brennan, 1989). Models and methods of G theory are concerned with the sources of error as separate entities from these multiple sources. A G theory approach to reliability is presented in the empirical evidence section at the end of this paper. Classical reliability estimates such as Cronbach's Alpha, equivalency, and stability coefficients are computed using the Analysis of Variance mecomponents. Establishing reliability and some form of rater agreement is critical for any classroom observation procedure used for licensing or certifying teachers.

Simplified Interrater Indices

Variance component calculations require an extensive knowledge of statistics and measurement. Ususally individuals who are involved with classroom observers, do not have the background to do the necessary calculations for determining reliability. Interrater reliability is calculated using a Pearson Product Moment Correlation and may even be too ambitious a project for some individuals. The authors of this paper propose a less sophisticated approach for establishing the degree of agreement or lack of agreement between two raters. A rater difference index (RDI) can be calculated for rating teaching behavior having k points or intervals using the following formula. The maximum value is 1 and the minimum value is 0. The closer the results are to 0 the better the RDI index rating is. For ratings, R_1 and R_2 obtained for two raters ratings the same behavior, the

$$\text{Rater Difference Index} = \frac{\left| R_1 - R_2 \right|}{(k - 1)} .$$

The rater difference index for the total scale score or instrument total score can be calculated by summing the differences for each teaching behavior and dividing by k-1 times the total number of behavior (items) in the analysis.

$$\text{Total RDI} = \frac{\Sigma \left| R_1 - R_2 \right|}{(k-1)\ (\text{number or items})} .$$

The same equations can be used to compute an absolute rater (ARI) and an absolute rater total index (ARTI) by using a norm value for a particular behavior in the

denominator for the ARI and the total norm value times the total number of behaviors for the TARI. The state of Florida uses a frequency count as an outcome measure for each behavior and has norm data available to raters. These indices would enable us to calculate a ratio of the difference between raters to a norm value.

Validity of Classroom Observations for Licensure and Certification

The first feature in this paper states that "The observation inference is established apriori and confirmed post hoc." This feature is stated in this manner because it is impossible to design a classroom observation that is going to be used for licensure or certification purposes without knowing what inference(s) is to be made regarding the outcome. The inference should be the foundation upon which the instrumentation is designed. Table 3 is four-fold classification featuring the meaning and values in test interpretation and test used (Messick, 1989). The four cells correspond, in general, to four inter-related features of fundamental validity which stand for the empirical foundation of score interpretation, meaning, basis for test use, the value implications, and social consequences. Messick incorporates construct validity in each cell thus binding "validity of test use to the validity of test interpretation... (1989, p.10)." Validity is critical to all aspects of the processes presented in this paper; it is the linchpin holding the entire process together.

Table 3. Facets of Validity as a Progressive Matrix

TABLE 3
Facets of Validity as a Progressive Matrix

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

Messick (1989, p.10)

Validity of Classroom Observations for Nonlicensure and Noncertification

Context observations of classroom teachers represent a more advanced methodology than previous checklists, observation systems, and rating scales. Recognition of the complexity of the teaching - learning process makes it essential that observations of classroom teachers include information about the context in which the observation is made. Details that include time-of-day, time-of-year, weather conditions, events preceding and scheduled events to follow the observation are important aspects of what happens in a classroom. The observation cannot be accurate or useful if it is no more than a snapshot of behaviors covering only a small portion of the continuous flow that constitutes the classroom.

Evaluation of teachers through use of context information and the model suggested in this paper applies best to the competent or better teacher, and possibly less well to the beginning, inexperienced, or less than competent teacher. More structure and a more consistent format are necessary for observation plans developed for teachers that are less than competent or inexperienced. The observation of the competent level or better teacher should be more on the order of peer collaboration or a type of self-initiated evaluation with peer assistance. In the observation system for competent teachers only the first four steps in the observation schemata, observe, record, retrieve, and analyze are necessary. Judging in step five, would not be considered because as no judgment would be rendered except by the teacher in assessing his/her own performance with feedback from the observer. If observation of competent teachers is the focus of

the procedure described in this paper, what criteria should be used as the basis for such a procedure? More to the point, how we insure that these observations are valid?

The first consideration should be the qualifications and background of the observer. This person should be at least as competent and experienced as the teacher who is being observed. A person who has taught the second grade for fifteen years should be observed by someone who has had a similar background and experience; that is someone who knows the context. Subject matter teachers in high school should be observed by the same type of teacher. Chemistry teachers should be observed by chemistry teachers with like background and experience. One of the problems with observations as we presently employ them is the use of principals, assistant principals, or supervisors who may have been art teachers, or mechanical drawing instructors, or the like in observing a large variety of teachers in situations for which that they have no inside knowledge based on experience. When checklists and rating scales are the basis for observations, training sessions are required of observers to insure that they are familiar with them and competent in their use. The training that an individual has received in a context observation system is not some external short-term indoctrination session, but rather it is the sum-total of their experience.

Using competent teachers to observe competent teachers and using their mutual background as a basis for pairing emphasizes the validity of the observation that is made and departs from the more traditional procedures that emphasize the use of objective instruments and training of observers to establish reliability. The type of observation procedure advocated by this paper emphasizes the validity of the observation from the standpoint of the expertise of the observer and the congruence of background between

the observee and the observer. This is closely related to qualitative research methods in which an ethnographer or qualitative researcher must gather sufficient information so as to know how the subject of research (in this case the observed teacher) feels and thinks. The old indian saying "Walk a mile in my moccasins" exemplifies this relationship between the observer and the observee. Elliot Eisner (1985) has referred to this notion as educational "connoisseurship" and criticism in which the background and experience of the observer or critic is emphsized.

A second area of emphasis is establishing the validity of context observations is recording. Observations should be recorded in as much detail as possible keeping in mind the context of behaviors and the importance that context has for analyzing the data later. Field notes recorded by qualitative researchers are something closely related to the recording methods suggested here. It is obvious to anyone familiar with school classrooms that recording everything is an impossible task, and the solution to the problem comes in narrowing the focus of the observation. This narrowing of the focus will make the task more manageable. The focus should be determined by the observed teacher, since it should be his/her request that initiates the observation. Competent teachers do not need observers to record routine behaviors that may be a necessity for beginning, inexperienced, or less competent teachers. The observation should focus on specific problems or questions that a teacher may have. For example, a teacher may request an observation to focus on the sequencing of content in a science lesson, or a teacher may request an observation to focus on small group work in social studies. Two points are important here. First, recorded observations should be focused and include

detailed context information. Second, the observation should be made at the request of the observed teacher as a part of professional growth through peer collaboration.

A final area of consideration for establishing the validity of context observations is analysis or use of the recorded observations. The third and fourth steps in the observation schemata are retrieve and analyze and should be accomplished in a joint effort between the observer and observee. Expanding and transcribing the observation notes should be done cooperatively by the two teachers involved with each contributing to the final draft of the recorded observations. Analysis should involve qualitative procedures that emphasize looking for common elements in the data. This emphasis on a more holistic examination of the data is quite different from the usual quantitative where analysis (When checklists and rating scales are used) involving, at the very least, frequencies and percentages. That is not to say that some quantitative methods cannot be used, but the emphasis should be on a larger picture of the classroom as a whole with context considered an essential integral part of the whole.

Emphasis on validity for context observations defined as usefulness or meaningfulness, is based primarily on the methods and procedures employed by qualitative researchers. The definition of valididy is of an emphasis of accuracy and comprehensiveness of the data involved than it is a dependency on quantitative definitions. Reliability is not an important consideration, since te validity is emphasized through the choice of observer, the contextual detail of the observations, and the cooperative effort involved in the retrieval and analysis of the data. Insuring that the data reflect, as much as possible, the reality of the classroom observed is of primary

importance. Knowing that two observers can agree is not considered as important as knowing that what was observed and recorded was an accurate portrayal of the situation.

Reliability and Validity of Context Classroom Observations

An Empirical Reliability and Validity of a Classroom Observation Instrument

Data were collected from the Kansas Internship Program using a Classroom Behavior Activity Record (CBAR). The CBAR is designed to capture the classroom context and accurately record the events of the classroom. The CBAR is used with a focused observation system adaptation from the Classroom Activity Record (CAR) Evertson, 1987). Use of the CBAR allows for classroom observation data to be collected capturing the subtle instructional and contextual features of a classroom in a standardized manner for research as well as evaluation purposes. The CBAR provides a systematic structure for documenting the occurrence of classroom behaviors and recording classroom activities. A series of codes denote a quick reference to the classroom activity which capture the instructional and managerial context of the classroom and the teaching behavior is indicated by the behavior code (Evertson and Burry, 1989). The behavior code refers to specific behaviors in the Kansas Internship Assessment Inventory (Poggio, Burry, and Glasnapp, 1988). Space is provided for relevant descriptive notes to be recorded simultaneously with the activity and behavior codes. Information regarding subject matter, grade level, school district, school, teacher, observer, date, and lesson are recorded at the top of the CBAR.

The CBAR may be implemented in any type of classroom with teachers at any level. It can be used at all experiential levels with a variety of observation systems and

including those which may also require anecdotal records, critical incident recordings, specimen descriptions, and on-line checklists of relevant behaviors.

Figure 1 illustrates the CBAR used to observe an assessment set of behaviors from the Kansas Internship Assessment Inventory.

The demographic data are located at the top and the behaviors to be observed during a focused classroom observation follow. Only information pertinent to the behavioral statements in the assessment set is to be recorded, thereby focusing the observer's attention on the in-depth details for five or six behaviors. The behavior code, activity code and descriptive notes follow. This is where the information for the observe and record steps are to be completed. The continuation of this CBAR (Figure 2) was designed to complete the retrieval, analysis/synthesis and judgement steps. It can be easily seen that the observation form was created to implement the five step observation schemata and facilitate the essential components of higher level processes.

Insert figures 1 and 2 about here

Study on Contextual Classroom Observation Data

CBARS were randomly selected from the 1987-88 Kansas Internship Program pilot study from 5 assistance committees (5 administrators and 5 matched senior teachers). Descriptive notes were analyzed for accuracy relevant to the behavioral definition for each of the 5 behaviors in the analysis. Figure 3 is an illustration of how one behavior is defined. Table 4 details the 5 behaviors selected for analysis by pilot study

observation groups. Six matched pairs were randomly selected for each assessment set of behaviors.

Table 4. Behaviors Used in the Analysis from the Kansas Internship Assessment Inventory

1.	Reflects an understanding of learning theory in planning and instruction.
2.	Is careful to focus student attention on important points in class lessons.
3.	Encourages questions and discussions from all students using effective questioning patterns and techniques.
4.	Reteaches concepts/skills students are not learning.
5.	Summarizes or achieves closure.

CBARS were analyzed by the researcher for accuracy and quality of the observers' descriptive notes. Ratings of "0" and "1" were given to each of the descriptive notes and percentages were obtained for the accuracy of each behavioral statement for administrators and senior teachers.

Reliability and validity indices were obtained from and analysis of variance nested design calculated using the general linear model (GLM) procedure from the Statistical Analysis Systems (SAS). Subjects were nested under raters and type III sums of squares and mean estimates were obtained for raters, behaviors, interaction between behaviors and raters, and the error term for subjects times behaviors nested under raters. Type III sums of squares for a particular variable is the sums of squares due to adding that variable to a model that already contains all the variables that precede the particular variable in the model statements. In this case type I and type III are the same, but type

III sums of squares are normally used to make inferences (Freund, Littell, & Spector, 1986). Table 6 details the ANOVA table for the accuracy ratings of the administrator and teacher ratings.

Table 5. Analysis of Variance Nested Design for the Accuracy of Adminstrator and Teacher Ratings of Behaviors

SOURCE	DF	TYPE III SS	F VALUE	PR > F
Rater	1	38.720	0.02	0.8978
Subj (Rater)	8	17614.560	5.21	0.0100
Measure	4	3357.08	1.99	0.1201
Rater*Behavior	4	2142.28	1.27	0.3028
Sub*Behav(Rater)	32	13512.24		

The appropriate error term for all of the F statistics, except rater times behavior, is the error SS obtained for the subject times behavior nested under raters. The rater time behavior interaction was used to obtain the F statistic for the rater SS. Reliability estimates were obtained using the Hoyt ANOVA Reliability estimate (Felt & Brennan, 1989; Crocker & Algina, 1986). The Hoyt ANOVA Reliability estimate is algebraically identical to Cronbach alpha and is the foundation upon which generalizability theory is derived. The formulas for Hoyt and the Generalizability coefficient are also identical.

$$\rho_{xx'} = \frac{MS_{(p)} - MS_{(r)}}{MS_{(p)}} = \frac{2201.82 - 422.2575}{2201.82} = .808$$

Another formula which makes use of adding both of the ANOVA error terms to the mean square persons in the denominator leaving the numerator the same, reveals a

reliability coefficient of .866 (Kane, 1986). Both coefficients reveal that the accuracy of the observations taken from the context instrument is reliable.

A validity check for the accuracy of the descriptive notes is addressed in two ways and two basic validity questions were addressed. First, is there a difference between administrators and senior teachers on the quality and accuracy of the descriptive notes? Second is there a difference in the quality and accuracy of information collected on the descriptive notes across five behaviors? A validity coefficient was obtained (Kane, 1986) using the mean squares from the ANOVA table.

$$\rho^2 = \frac{MS_{(p)}}{MS_{(p)} + MS_{(rm)} + MS_{(e)}} = \frac{2201.82}{2201.82 + 535.57 + 422.26} = .8347$$

All of the F statistics from the ANOVA are not significant which infers that there is no difference in the accuracy of the descriptive notes across administrators and behaviors.

According to the Standards on Psychological Testing (APA, AREA, NCME, 1984) validity should be established on the use of the instrument. The accuracy of the descriptive notes are the vehicle on which a judgment is made and are a critical element of the test use. The answers to these questions document the validity of the use of the Classroom Behavior Activity Record as a measure of the classroom context.

Earlier in this paper validity was addressed. As Messick has emphasized (1989)

> "Validity is an integrated valuative judgment of the degree to which empirical evidence and theoretic rationales support the adequacy and appropriateness of inferences and actions based on test scores of other modes of assessment" (p. 13).

Using this definition and the empirical data described here it can be inferred that there is a degree of cutting validity evidence for the accuracy of the use of the Classroom Observation System.

Conclusions

Reliability and validity are the essence of the assessment process and practice. Instruments must undergo a comprehensive analysis for appropriate psychometric properties before and after they are implemented. Classroom observations must be subjected to the same rigorous process as any other standardized testing procedure; however, the trial has just begun for determing whether our traditional conventions work.

Messick's statements regarding validity prompt us that it is critical to maintain the integrity of validity. We need to build our instruments and procedures to support the specific uses and inferences to be made from the scores. Seven features for valid classroom assessment are presented in this paper in attempt to focus and guide on the specific uses of classroom observations. The features are: the observation inference is established apriori and confirmed post hoc (craft knowledge), expertise in pedagogy, is developmental; instrumentation must be built upon a research base, empirical base, or a combination of the two; observation criteria, when, appropriate, are both generic and content specific; observers are well trained in the process of doing an observation; observers are well-honed on the meaning(s) of the teaching behaviors to be observed; and procedures must vary depending upon the inferences(s) that are made. If all seven features are present then the observation is valid.

Reliability and validity evidence support the accuracy of the use for a somewhat unique classroom context observation instrument used with the Kansas Internship Assessment Inventory entitlted the Classroom Behavior Activity Recored (CBAR). The CBAR, because of its qualitative nature, is flexable and appropriate for use with any experiential level, teaching context, teaching content or grade level. Hoyt reliability coefficients are above .80 for accuracy ratings and valididty is demonstrated by Kane's validity coefficeint of .86 and the nonsignigicant F statistic computed from the nested analysis of variance.

The trial has begun, but it is a very long way from a verdict. Many more studies need to be conducted on our present and future instruments. A verdict needs to be awarded for each specific use to be made of a classroom observation istrument as was illustrated in this paper.

The paradox of reliability and validity should not exist. Are teachers not entitlted to the most valid procedures and instruments we can design to assist them in their professional growth and to obtain licenses and certifications? Let us continue our efforts to further inevestigate the psychometric properties that will enable us to use these instruments to guide our teachers into being the best possible educators.

References

Allen, M. & Yen, W. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Ary, Jacobs, & Razavich. (1990). Introduction to research in education. Fort Worth, TX: Holt Reinhard and Winston.

Berliner, D. (1988). Implications of studies of expertise in pedagogy for teacher education and evaluation. New Directions for teacher Assessment. Princeton, New Jersey: Educational Testing Service, 39-68.

Bolton. D. (1972). Selection and Evaluation of Teachers. Berkley, CA: Mc Cutchan.

Brophy, J. & Good,T. (1974). Teacher-student relationships; Causes and Consequences. New York: Holt, Rinehart and Winston.

Burry, J., Chissom, B. & Shaw, D. (1990 April, 1990). Validity and reliability of context teacher performance assessment. Paper presented at the National Council in Measurement and Education, Boston, MS.

Burry, J., Poggio, J., & Glasnapp, D. (1989). The Kansas internship program assistance/assessment model. Journal of Personnel Evaluation in Education .

Burry, J., & Shaw, D. (1988). Teachers and administrators differ in assessing teacher effectiveness. Journal of Personnel and Evaluation in Education, 2:33-41, 33-41.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt Reinhart and Winston.

Educational Research Service (1978). Evaluating teacher performance. Arlington, VA: Author.

Eisner, E. (1985). The art of educational evaluation a personal view. Great Britain: The Falmer Press.

Evertson, C. (1987). Classroom Activity Record: Observation record for Project STAR. Nashville, TN: Vanderbilt Institute for Public Policy Studies, Center for Education Policy, Vanderbilt University.

Felt, L. & Brennan, R. (1989). Reliability. In M. Wittrock (Ed.), Handbook of Research on Teaching (3rd ed.) New York:

Freund, R., Littell, R., & Spector, P. (1986). Sas Systems for linear models. Cary, NC: Sas Institute Inc.

Joint Committee on Standards for Educational Evaluation. In D. Stuffelbeam, (Ed.), The Personnel evaluation standards: How to assess systems for educators. Newbury Park, CA: Sage Publications, Inc.

Kane, M. (1982). A sampling model for validity. Applied Psychological Measurement, 6, 187-200.

Landy, F., & Farr, J. (1983). The Measure of Work Performance. New York: Academic Press.

Leinhardt, Gaea. (1990). Capturing Craft Knowledge in Teaching. Educational Researcher. Vol. 19 No. 2.

Magnusson, (1966). Test theory. Stockholm: Addison-Wesley.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. Educational Researcher, 18, pp. 5-11.

Messick, S. (1989). Validity. In M. Wittrock (Ed.), Handbook of Research on Teaching (3rd ed.) New York:

Poggio, J., Glasnapp, D., & Burry, J. (1989, March). Teaching characteristics a search for classification. Paper presented at the National Convention for the National Council in Measurement in Education, San Francisco, CA.

Poggio, J., Burry, J., & Glasnapp, D. (1988). Kansas Internship Assessment Inventory. Lawrence, KS: Center for Educational Testing and Evaluation at the University of Kansas.

Rosenshine, B. (1971). Teaching Behaviros and Student Achievement. London: National Foundation for Educational Research.

Shavelson, R., Webb, N., & Burstein, L. (1986). Measurement of teaching. In M. Wittrock (Ed.), Handbook of Research on Teaching (3rd ed.) New York: Macmillan, 50-91.

Shaw, D., Huffman, M., & Haviland, M. (197) Grouping continuous data in discrete intervals: information loss and recovery. Journal of Educational Measurement, 24, 167-173.

Shulman, L. Bird, T. & Haertel, E. (April, 1988). Toward alternative assessments of teaching: A report of work in progress. Stanford, CA: Teacher Assessment Project, Stanford University School of Education.